AD-A056 705 STATE UNIV OF NEW YORK AT BUFFALO AMHERST STATISTICA--ETC F/G 12/1

TECHNIQUES OF QUANTILE REGRESSION.(U)

JUN 78 J CARMICHAEL DAAG29-76-G-0239

UNCLASSIFIED ARO-13845.3-M NL

1 OF 1

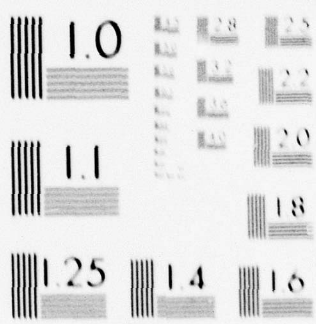
AD A056 705

END

DATE FILMED

9-78

DDC



MICROCOPY RESOLUTION TEST CHART

NATIONAL BUREAU OF STANDARDS-1963-A

State University of New York at Buffalo
Department of Computer Science

LEVEL II

18 ARO 13845.3-M 19

12

6 # TECHNIQUES OF QUANTILE REGRESSION*

9 Technical rept.,

by

10 Jean-Pierre Carmichael

Statistical Science Division
State University of New York at Buffalo

15 DAAG29-76-G-0239

~~GRANT TECHNICAL REPORT NO. ARO-5~~

11 Jun [redacted]78

12 27p.

*Research supported by the Army Research Office (Grant DA AG29-76-0239).

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER ~~Technical Report No. ARO 5~~	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) Techniques of Quantile Regression		5. TYPE OF REPORT & PERIOD COVERED Technical
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Jean-Pierre Carmichael		8. CONTRACT OR GRANT NUMBER(s) DA AG29-76-G-0239
9. PERFORMING ORGANIZATION NAME AND ADDRESS Statistical Science Division State University of New York at Buffalo Amherst, New York 14226		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
11. CONTROLLING OFFICE NAME AND ADDRESS U. S. Army Research Office Post Office Box 12211 Research Triangle Park, NC 27709		12. REPORT DATE June 1978
		13. NUMBER OF PAGES 27
14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

NA

18. SUPPLEMENTARY NOTES

The findings in this report are not to be construed as an official Department of the Army position, unless so designated by other authorized documents.

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Density Estimation
Nonparametric Regression
Order Statistics

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

We study the asymptotic properties of different nonparametric estimators of a regression function. are studied The motivation for these estimators comes from the unified approach to statistics developed by Parzen under the name of "Nonparametric Statistical Data Science" in which the quantile function plays a crucial role. We call them quantile regression estimators.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-014-6601

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

78 07 21 028

TECHNIQUES OF QUANTILE REGRESSION*

by

Jean-Pierre Carmichael

Introduction

Given observations $\{(X_i, Y_i), i = 1, \ldots, n\}$ on random variables (X, Y) with joint distribution $F_{X,Y}(x, y)$, we want to estimate the regression function of Y on X, $E[Y | X = x]$, nonparametrically.

In order to find a natural estimator (simple computationally and intuitively appealing), Parzen (1977) developed the following theoretical approach.

1. Theoretical Approach:

Let $U_1 = F_X(X)$ and $U_2 = F_Y(Y)$, then the joint distribution of U_1 and U_2 is

$$D_{U_1, U_2}(u_1, u_2) = F_{X,Y}\left(Q_X(u_1), Q_Y(u_2)\right)$$

and their joint density is

$$d_{U_1, U_2}(u_1, u_2) = \frac{f_{X,Y}\left(Q_X(u_1), Q_Y(u_2)\right)}{f_X\left(Q_X(\mu_1)\right) f_Y\left(Q_Y(u_2)\right)}$$

where F_Z is the distribution function of Z

f_Z is its density function

Q_Z is its quantile function

---

*Research supported by Army Research Office (Grant DA AG29-76-0239).

Let $r(x)$ be the regression function of Y on $X = x$.

$$r(x) = E[Y | X = x] = \int_{-\infty}^{\infty} \frac{y\, f_{X,Y}(x, y)\, dy}{f_X(x)}$$

We now define the regression-quantile function $rQ(\cdot)$ by

$$rQ(u) = r\left(Q_X(u)\right) = E[Y | X = Q_X(u)]$$

How do we compute $rQ(\cdot)$?

By definition,

$$rQ(u) = \int_{-\infty}^{\infty} \frac{y\, f_{X,Y}\left(Q_X(u), y\, dy\right)}{f_X\left(Q_X(u)\right)}$$

Let $y = Q_Y(u_2)$, then

$$rQ(u) = \int_0^1 Q_Y(u_2)\, d_{U_1, U_2}(u, u_2)\, du_2$$

If we introduce a Dirac delta function, we can express $rQ(\cdot)$ as a double integral

1.1 $$rQ(u) = \int_0^1 \int_0^1 Q_Y(u_2)\, \delta(u_1 - u)\, d\, D_{U_1, U_2}(u_1, u_2)$$

We estimate $rQ(\cdot)$ by

1.2 $$\hat{rQ}(u) = \int_0^1 \int_0^1 \hat{Q}_Y(u_2) \frac{1}{h(n)}\, K \left(\frac{u_1 - u}{h(n)}\right) d\, \hat{D}_{U_1, U_2}(u_1, u_2) \quad .$$

$\hat{D}_{U_1, U_2}(\cdot\ ,\ \cdot)$ is an estimator of the joint distribution function of U_1 and U_2. It could be the empirical joint distribution function.

$\hat{Q}_Y(\cdot)$ is an estimator of the quantile function of Y. It could be the empirical quantile function of the Y's.

$K(\cdot)$ is an approximator to the Dirac delta function.

2. Different Estimators:

Let $Y_{[i\,:\,n]}$ be the observation associated with $X_{(i)}$, where $X_{(1)} < X_{(2)} < \ldots < X_{(n)}$. $Y_{[i\,:\,n]}$ is called the concomitant of the i^{th} order statistic.

Let $\hat{D}_{U_1, U_2}(\cdot\ ,\ \cdot)$ be the empirical joint distribution function of U_1 and U_2. It has jumps of size $1/n$ at points of the form $(i/n, R_i/n)$ where R_i is the rank of $Y_{[i\,:\,n]}$ among the Y's.

Let $K(\cdot)$ be a kernel function with bandwidth parameter $h(n)$.

A first estimator of $\hat{Q}_Y(\cdot)$ is given by

$$\hat{Q}_1(u_2) = Y_{[i\,:\,n]}\ , \quad \frac{i-1}{n} \le u_2 < \frac{i}{n}$$

Then, equation 1.2 becomes

$$2.1 \qquad r\hat{Q}_1(u) = \sum_{j=1}^{n} Y_{[j:n]} \int_{\frac{j-1}{n}}^{j/n} \frac{1}{h(n)}\ K\left(\frac{t-u}{h(n)}\right)dt$$

Usually, as in Yang (1977), this is approximated by

2.2 $$r\hat{Q}_{(1)}(u) = \frac{1}{u} \sum_{j=1}^{n} Y_{[j\,:\,n]} \; K\left(\frac{j/n - u}{h(n)}\right) \cdot \frac{1}{h(n)}$$

Yang studied the statistical properties of that estimator.

Note that $r\hat{Q}_{(1)}(\cdot)$ can be viewed as the result of smoothing amplitudes $Y_{[j\,:\,n]}$ observed at equidistant points of the form j/n.

Clark (1977) recommends to interpolate linearly between the successive points $\{(j/n, Y_{[j\,:\,n]})\}$ to get an estimator with maybe more derivatives than the kernel.

Define

$$\hat{Q}_2(u_2) = \begin{cases} Y_{[1\,:\,n]}\,, & 0 \le u_2 \le 1/n \\ Y_{[j\,:\,n]}(j + 1 - nu_2) + Y_{[j+1\,:\,n]}(nu_2 - j)\,, & \\ & \frac{j}{n} \le u_2 \le \frac{j+1}{n}\,, \quad j = 1, \ldots, n-1 \end{cases}$$

Then

2.3 $$r\hat{Q}_2(u) = \int_0^1 \hat{Q}_2(t)\,\frac{1}{h(n)}\,K\left(\frac{t-u}{h(n)}\right)dt\,.$$

It has been remarked before that it is difficult to smooth a curve unless it is relatively flat. People would then recommend to subtract a trend term from the data before smoothing.

We would like to propose instead to smooth the first differences $\hat{Q}'_2(\cdot)$,

$$\hat{Q}'_2\ (u_2) = \begin{cases} 0 & 0 \le u_2 < 1/n \\ n \cdot \left(Y_{[j+1:n]} - Y_{[j:n]}\right), & \frac{j}{n} \le u_2 < \frac{j+1}{n} \end{cases}$$

$$j = 1, \ldots, n-1$$

We then form the estimator $r\hat{Q}'(\cdot)$

2.4 $$r\hat{Q}'_1(u) = \int_0^1 \hat{Q}'_2(t) \frac{1}{h(n)} K\left(\frac{t-u}{h(n)}\right) dt$$

$$r\hat{Q}'_1(u) = \sum_{j=1}^{n-1} n\left(Y_{[j+1:n]} - Y_{[j:n]}\right) \frac{1}{h(n)} \int_{j/n}^{j+1/n} K\left(\frac{t-u}{h(n)}\right) dt$$

and

2.5 $$r\hat{Q}_3(u) = \int_{1/2}^{u} r\hat{Q}'(s)\, ds + r\hat{Q}_1(1/2)$$

Because an estimator of $r\hat{Q}(\cdot)$ would be the indefinite integral of $r\hat{Q}'(\cdot)$, we fix the value of $r\hat{Q}_3(1/2)$ to be $r\hat{Q}_1(1/2)$ as we feel that all estimators are usually good for the middle values. The problems and the differences between estimators usually appear near the endpoints.

Finally, we can smooth $\hat{Q}'_2(\cdot)$ using the autoregressive method by computing its Fourier coefficients

$$\hat{\varphi}(v) = \int_0^1 e^{2\pi i t v}\, \hat{Q}'_2(t)\, dt$$

2.6 $$\hat{\varphi}(v) = \sum_{j=1}^{n-1} n\left(Y_{[j+1:n]} - Y_{[j:n]}\right) \int_{j/n}^{j+1/n} e^{2\pi i t v}\, dt$$

$$|v| = 0, 1, 2, \ldots, m$$

From the $\hat{\varphi}(\cdot)$'s , we compute the autoregressive coefficients by solving the Yule-Walker equations

2.7 $$\hat{rQ_2'}(u) = \frac{\hat{\sigma}_k^2}{\left|1 + \sum_{j=1}^{k} \hat{\alpha}_{j,k} e^{2\pi iju}\right|^2}$$

and

2.8 $$\hat{rQ_4} = \int_{1/2}^{u} \hat{rQ_2'}(s)\, ds + \hat{rQ_1}(1/2) \ .$$

Note the relation between the linearized version of the data and first differences. Taking k^{th} order differences would be like interpolating between data points with a k^{th} degree polynomial.

3. Statistical Properties:

3.1 General Results

Yang (1977) studied statistical properties of linear functions of concomitant of order statistics.

Among the different estimators proposed in the previous section, only $\hat{rQ_2'}(\cdot)$ and $\hat{rQ_4}(\cdot)$ are not of that form.

For convenience, we reexpress the three major results of Yang in a form more related to our purpose.

We need the following:

Let $$M_n = \int_0^1 \int_0^1 \hat{g}(u_2) \frac{1}{h(n)} K\left(\frac{u_1 - u}{h(n)}\right) d\,\hat{D}_{U_1, U_2}(u_1, u_2)$$

where $$\hat{g}(u_2) = H\left(X_{(i)}, Y_{[i:n]}\right), \quad \frac{i-1}{n} \le u_2 < \frac{i}{n} .$$

$$\alpha(x) = E[H(X, Y) \mid X = x]$$

$$\sigma^2(x) = \mathrm{Var}\,(H(X, Y) \mid X = x)$$

Assumptions

A1 - $E[\,|H(X, Y)|^2\,] < \infty$

A2 - $\alpha(x)$ can be expressed as a difference of two increasing right-continuous functions

A3 - $\sigma^2(x)$ has the same property as $\alpha(x)$ or $F_X(x)$ is absolutely continuous

A4 - $\alpha\left(Q(t)\right)$ is continuous at $t = u$

A5 - $E[H(X, Y)^3] < \infty$

A6 - $\alpha'(x) = \frac{d}{dx}\alpha(x)$ exists and $\alpha'\left(Q(t)\right)$ is continuous at $t = u$, $0 < u < 1$

A7 - $\frac{d^2}{dt^2}\alpha\left(Q(t)\right)$ exists and is continuous at $t = u$

B1 - There exists $M > 0$ such that

$$|K(t_1) - K(t_2)| < M \cdot |t_1 - t_2| \quad \text{for all } t_1, t_2$$

B2 - $|tK(t)| \to 0$ as $|t| \to 1$

B3 - $\int_{-1}^{1} K(t)\,dt = 1$

B4 - $\lim\limits_{n\to\infty} h(n) = 0$

B5 - $\lim\limits_{n\to\infty} h^{-1}(n)\left(\dfrac{\log\log n}{n}\right)^{1/4} = 0$

B6 - $\int_{-1}^{1} tK(t)\,dt = 0$

B7 - $\int_{-1}^{1} t^2 |K(t)|\,dt < \infty$

B8 - $K''(t)$ exists and satisfies B1 and B2

$\text{Th}^{\text{m}}1$ - Consistency (Yang's Theorem 5)

Under assumptions A1 - A4 and B1 - B5,

$$\lim_{n\to\infty} E[M_n] = \alpha\left(Q(u)\right)$$

$$\lim_{n\to\infty} E[M_n] = \lim_{n\to\infty} \int_0^1 \alpha\left(Q(u_1)\right) \cdot \frac{1}{h(n)} K\left(\frac{u_1 - u}{h(n)}\right) du_1$$

$$= \alpha\left(Q(u)\right) \cdot \int_{-1}^{1} K(t)\,dt = \alpha\left(Q(u)\right)$$

and

$$\lim_{n\to\infty} E[|M_n - \alpha\left(Q(u)\right)|^2] = 0$$

$\text{Th}^{\text{m}}2$ - Asymptotic normality (Yang's Theorem 6)

Under assumptions A1 - A6 and B1 - B5 ,

$$\sqrt{nh(n)}\ (M_n - E[M_n]) \xrightarrow{D} N\left(0,\ \sigma^2\left(Q(u)\right) \int_{-1}^{1} K^2(t)\,dt\right)$$

$\text{Th}^{\text{m}}3$ - Asymptotic bias (Yang's Corollary 1 to Theorem 6)

Under assumptions A1 - A7 and B1 - B8 ,

$$\lim_{n\to\infty} \frac{E[M_n] - \alpha\left(Q(u)\right)}{h^2(n)} = \frac{d^2}{du^2}\,\alpha\left(Q(u)\right) \cdot \int_{-1}^{1} t^2 K(t)\,dt$$

and $$\sqrt{nh(n)}\ \left(M_n - \alpha\left(Q(u)\right)\right) \xrightarrow{D} N\left(0,\ \sigma^2\left(Q(u)\right) \cdot \int_{-1}^{1} K^2(t)\,dt\right)$$

Let us apply these general results to the different estimators we presented in the previous section.

3.2 Statistical Properties of $\hat{rQ}_1(\cdot)$ and $\hat{rQ}_{(1)}(\cdot)$

$\hat{rQ}_{(1)}(\cdot)$ is the estimator proposed and studied explicitly by Yang as an estimator of $E[Y \mid X = Q(\cdot)]$. Our $\hat{rQ}_1(\cdot)$ has exactly the same properties as can be seen from the fact that

$$\int_{\frac{j-1}{n}}^{j/n} \frac{1}{h(n)}\,K\left(\frac{t-u}{h(n)}\right)dt = \frac{1}{nh(n)}\,K\left(\frac{t_j^* - u}{h(n)}\right)$$

$$\text{for } \frac{j-1}{n} \le t_j^* \le j/n$$

Thus, $\hat{rQ}_1(u)$ is a consistent estimator of $rQ(u) = E[Y|X = Q(u)]$, under the conditions of Theorem 1, at the points of continuity of $rQ(\cdot)$.

Under the conditions of Theorem 3, the asymptotic bias is proportional to the second derivative of $rQ(\cdot)$

For the kernel we have been using

$$K(z) = \begin{cases} \frac{15}{16}(1 - z^2)^2 & |z| \le 1 \\ 0 & |z| > 1 \end{cases}$$

the asymptotic bias is $1/7 \cdot rQ''(u)$ and the variance of the asymptotic distribution is $5/7 \cdot \mathrm{Var}\left(Y|X = Q(u)\right)$.

It is possible to estimate $\mathrm{Var}\left(Y|X = Q(u)\right)$ by the same method, e.g.

$$\hat{\sigma}^2\left(Y|X = Q(u)\right) = \frac{1}{n} \sum_{j=1}^{n} \left(Y_{[j:n]} - \hat{rQ}_{(1)}(u)\right)^2 \frac{1}{h(n)} K\left(\frac{j/n - u}{h(n)}\right) .$$

3.3 Statistical Properties of $\hat{rQ}_2(u)$

We rewrite $\hat{rQ}_2(\cdot)$ as follows:

$$\hat{rQ}_2(u) = I_1(u) + I_2(u) , \quad \text{where}$$

$$I_1(u) = \sum_{j=1}^{n} \int_{\frac{j-1}{n}}^{j/n} Y_{[j:n]} \cdot \frac{1}{h(n)} K\left(\frac{t-u}{h(n)}\right) dt$$

$$I_2(u) = \sum_{j=2}^{n} \int_{\frac{j-1}{n}}^{j/n} n \cdot \left(Y_{[j-1:n]} - Y_{[j:n]}\right)\left(\frac{j}{n} - t\right) \cdot \frac{1}{h(n)} K\left(\frac{t-u}{h(n)}\right) dt$$

Note that $I_1(u)$ is just $r\hat{Q}_1(u)$. On the other hand,

$$E[(Y_{[j-1:n]} - Y_{[j:n]})] = \int_0^1 rQ(s)\, d[\binom{n}{j-1} s^{j-1}(1-s)^{n-j+1}]$$

and by expanding in Taylor series

$$n\int_{\frac{j-1}{n}}^{j/n} (\frac{j}{n} - t) \cdot \frac{1}{h(n)} K(\frac{t-u}{h(n)})\, dt =$$

$$\frac{1}{2nh(n)} K\left(\frac{\frac{j-1}{n} - u}{h(n)}\right) + \frac{1}{n^2h^2(n)} \left[\frac{1}{2} K'\left(\frac{\frac{j-1}{n} - u}{h(n)}\right) - \frac{1}{3} K'\left(\frac{t_j - u}{h(n)}\right)\right] + R_{jn}$$

where $$|R_{jn}| \leq \frac{1}{6n^3h^3(n)} \left|K''\left(\frac{t_j - u}{h(n)}\right)\right|$$

$$\frac{j-1}{n} < t_j < \frac{j}{n}, \; j = 2, \ldots, n$$

The bias properties of $r\hat{Q}_2(\cdot)$ are the same as those of $r\hat{Q}_1(\cdot)$ provided $I_2(\cdot)$ contributes only to high order terms.

$$E[I_2(u)] = J_1(u) + J_2(u) + R$$

We look only at $J_1(\cdot)$.

$$J_1(u) = \int_0^1 rQ(s) \sum_{j=2}^{n} \left(2nh(n)\right)^{-1} K\left(\frac{\frac{j-1}{n} - u}{h(n)}\right) d\,[\,(\frac{n}{j-1})\, s^{j-1}(1-s)^{n-j+1}]$$

By Bernstein approximation,

$$J_1(u) = \left(2nh(n)\right)^{-1}\left[\int_0^1 rQ(s)h^{-1}(n)\,K'\left(\frac{s-u}{h(n)}\right)ds - \int_0^1 rQ(s)K\left(\frac{-u}{h(n)}\right)d(1-s)^n\right.$$

$$\left. - \int_0^1 rQ(s)\,K\left(\frac{1-u}{h(n)}\right)ds^n\right].$$

For $0 < u < 1$,

$$\int_0^1 rQ(s)\,h^{-1}(n)\,K'\left(\frac{s-u}{h(n)}\right)ds = \int_{u-h(n)}^{u+h(n)} rQ(s)h^{-1}(n)\,K'\left(\frac{s-u}{h(n)}\right)ds$$

because $K(\cdot)$ is defined only on $(-1,1)$ and upon integrating by parts, this is

$$h(n)\cdot\int_{-1}^{1} rQ'\left(u+th(n)\right)K(t)\,dt \doteq h(n)\,rQ'(u)$$

On the other hand,

$$\frac{1}{n}\left|\int_0^1 rQ(s)\,d(1-s)^n\right| < \left|\int_0^1 rQ(s)\,d(1-s)\right| = A$$

$$\frac{1}{n}\left|\int_0^1 rQ(s)\,d\,s^n\right| < \left|\int_0^1 rQ(s)\,ds\right| = B$$

and for $h(n) < \min(u, 1-u)$

$$\frac{A}{h(n)}\cdot K\left(\frac{-u}{h(n)}\right) = \frac{B}{h(n)}\cdot K\left(\frac{1-u}{h(n)}\right) = 0 \quad .$$

So,

$$J_1(u) = \left(2nh(n)\right)^{-1}\left(h(n)\ rQ'(u)\right)$$

goes to zero as n^{-1} . $J_2(\cdot)$ goes to zero faster.

Thus, for $0 < u < 1$,

$$\lim_{n\to\infty} \frac{E[r\hat{Q}_2(u) - rQ(u)]}{h^2(n)} = \lim_{n\to\infty} \frac{E[r\hat{Q}_1(u) - rQ(u)]}{h^2(n)} .$$

At the endpoints, the limit does not exist.

3.4 Statistical Properties of $r\hat{Q}_3(\cdot)$

We start by studying $r\hat{Q}_1'(\cdot)$.

$$r\hat{Q}_1'(u) = \sum_{j=2}^{n} n\left(Y_{[j:n]} - Y_{[j-1=n]}\right) \cdot h(n)^{-1} \int_{j-1/n}^{j/n} K\left(\frac{t-u}{h(n)}\right)dt .$$

By Taylor series expansion,

$$nh^{-1}(n) \int_{j-1/n}^{j/n} K\left(\frac{t-u}{h(n)}\right)dt = \frac{1}{h(n)}\ K\left(\frac{\frac{j-1}{n} - u}{h(n)}\right) +$$

$$\frac{1}{2nh^2(n)}\ K'\left(\frac{\frac{j-1}{n} - u}{h(n)}\right) + R$$

and

$$E\left[Y_{[j:n]} - Y_{[j-1=n]}\right] = -\int_0^1 rQ(s)\, d\left[\binom{n}{j-1} s^{j-1}(1-s)^{n-j+1}\right].$$

So,

$$E\left[\hat{rQ}_1'(u)\right] = -\int_0^1 rQ(s) \sum_{j=2}^{n} \left\{ h(n)^{-1} K\left(\frac{\frac{j-1}{n} - u}{h(n)}\right) + \frac{1}{2nh^2(n)} K'\left(\frac{\frac{j-1}{n} - u}{h(n)}\right)\right\}$$

$$d\left[\binom{n}{j-1} s^{j-1} (1-s)^{n-j+1}\right]$$

By Bernstein approximation,

$$E\left[\hat{rQ}_1'(u)\right] = -\int_0^1 rQ(s)\, h^{-2}(n)\, K'\left(\frac{s-u}{h(n)}\right) ds \; -\int_0^1 rQ(s) \frac{1}{2nh^3(n)} K''\left(\frac{s-u}{h(n)}\right) ds + R$$

For $0 < u < 1$,

$$E\left[\hat{rQ}_1'(u)\right] = -h^{-1}(n)\, rQ\big(u + th(n)\big)K(t)\Big|_{-1}^{1} + \int_{-1}^{1} rQ'\big(u + th(n)\big)\, K(t)\, dt + O(n^{-1})$$

Thus

$$E\left[\hat{rQ}_1'(u)\right] = rQ'(u)$$

and

$$\frac{E\left[\hat{rQ}_1'(u) - rQ'(u)\right]}{h^2(n)} \doteq \frac{rQ'''(u)}{2} \int t^2 \cdot K(t)\, dt \quad .$$

From these formulas, one can evaluate $E\left[\hat{rQ}_3(u)\right]$. The terms missing in the Bernstein approximation formula are zero if $h(n)$ is less than $\min(u, 1-u)$ as in the previous section. The integral involving $K''(\cdot)$ contributes a term of order $\left(nh^2(n)\right)^{-1}$ to the expected value. Its influence is not felt either in the bias $\left(\lim_{n\to\infty} nh^4(n) = \infty\right)$.

$$E\left[\hat{rQ}_3(u)\right] = \int_{1/2}^{u} E\left[\hat{rQ}_1'(s)\right] ds + E\left[\hat{rQ}_1(1/2)\right] = rQ(u)$$

and

$$\frac{E\left[\hat{rQ}_3(u)\right] - rQ(u)}{h^2(n)} = \int_{-1}^{1} t^2 K(t)\, dt \left[\frac{rQ''(u)}{2} + rQ''\left(\frac{1}{2}\right)\right]$$

4. Case of X fixed

We study only the case where the x's are fixed and equidistant on the unit interval, of the form $\{j/n\}_{j=0}^{n}$. The model is of the form $Y = f(x) + \epsilon$, where the ϵ's are uncorrelated errors with mean zero and constant variance.

We limit ourselves to only two estimators:

$$\hat{f}_1(u) = \frac{1}{nh(n)}\left[\frac{1}{2}\, Y(0) \cdot K\left(\frac{-u}{h(n)}\right) + \sum_{j=1}^{n-1} Y(j/n) \cdot K\left(\frac{j/n - u}{h(n)}\right) + \frac{1}{2} \cdot Y(1) \cdot K\left(\frac{1-u}{h(n)}\right)\right]$$

and the estimator based on first differences

$$\hat{f}_2(u) = \int_{1/2}^{u} \hat{f}_1{}'(s)\, ds + \hat{f}_1(1/2)$$

where $\hat{f}_1{}'(s) = \Sigma\left[Y\left(\frac{j+1}{n}\right) - Y(j/n)\right] \cdot \frac{1}{h(n)}\, K\left(\frac{\frac{j}{n} - s}{h(n)}\right)$

and $Y(j/n)$ is observed at $x = j/n$.

4.1 Statistical Properties of $\hat{f}_1(\cdot)$

$$E\left[\hat{f}_1(u)\right] = \frac{1}{nh(n)}\left[\frac{1}{2}\, f(0) K\left(\frac{-u}{h(n)}\right) + \sum_{j=1}^{n-1} f(j/n)\, K\left(\frac{j/n - u}{h(n)}\right) + \frac{1}{2}\, f(1) \cdot K\left(\frac{1-u}{h(n)}\right)\right].$$

This formula is recognized as the trapezoidal rule for $\int_0^1 f(t)\,\frac{1}{h(n)}\,K\left(\frac{t-u}{h(n)}\right)dt$ based on the given design, so

$$E\left[\hat{f}_1(u)\right] \underset{n\to\infty}{\longrightarrow} f(u)$$

As an approximation to the integral, the error is at most $\frac{1}{12n^2} \sup_{0\le u\le 1} f''(u)$, which is much less important than the error of approximation of the integral to $f(u)$ that was found before to be

$$h^2(n)\, f''(u) \int_{-1}^{1} t^2 K(t)\, dt \quad .$$

Thus, $$\frac{E\left[\hat{f}_1(u) - f(u)\right]}{h^2(n)} \to f''(u) \int_{-1}^{1} t^2\, K(t)\, dt \quad .$$

Because the ϵ's are uncorrelated,

$$\mathrm{Var}\left(\hat{f}_1(u)\right) = \frac{\sigma^2}{n^2 h^2(n)} \left[\frac{1}{4} K^2\left(\frac{-u}{h(n)}\right) + \sum_{j=1}^{n-1} K^2\left(\frac{j/n - u}{h(n)}\right) + \frac{1}{4} K^2\left(\frac{1-u}{h(n)}\right)\right] \quad .$$

Thus,

$$nh(n)\ \mathrm{Var}\left(\hat{f}_1(u)\right) \underset{n\to\infty}{\longrightarrow} \sigma^2 \int_{-1}^{1} K^2(t)\, dt \quad .$$

4.2 Statistical Properties of $\hat{f}_2(\cdot)$

$$E\left[\hat{f}_1'(s)\right] = \frac{1}{n}\sum_{j=0}^{n-1} \frac{f\left(\frac{j+1}{n}\right) - f\left(\frac{j}{n}\right)}{1/n} \cdot \frac{1}{h(n)} K\left(\frac{\frac{j}{n} - s}{h(n)}\right) \xrightarrow[n\to\infty]{} f'(s) .$$

The asymptotic bias is computed as in section 3

$$\frac{E\left[\hat{f}_1'(s) - f'(s)\right]}{h^2(n)} \xrightarrow[n\to\infty]{} \frac{f'''(s)}{2} \cdot \int_{-1}^{1} t^2 K(t)\, dt$$

Thus,

$$E\left[\hat{f}_2(u)\right] \to \int_{1/2}^{u} f'(s)\, ds + f(1/2) = f(u)$$

and

$$\frac{E\left[\hat{f}_2(u) - f(u)\right]}{h^2(n)} = \left(\frac{f''(u)}{2} + f''(1/2)\right) \cdot \int_{-1}^{1} t^2 K(t)\, dt$$

One can write an exact expression for the variance of $\hat{f}_2(\cdot)$:

$$\mathrm{Var}\left(\int_{1/2}^{u} \hat{f}_1'(s)\, ds + \hat{f}_1(1/2)\right) = \mathrm{Var}\left(\int_{1/2}^{u} \hat{f}_1'(s)\, ds\right) + \mathrm{Var}\left(\hat{f}_1(1/2)\right)$$
$$+ 2\,\mathrm{Cov}\left(\int_{1/2}^{u} \hat{f}_1'(s)\, ds\, ,\ \hat{f}_1(1/2)\right) .$$

Now, $\mathrm{Var}\left(\hat{f}_1(1/2)\right)$ was computed previously and

$$\mathrm{Var}\left(\int_{1/2}^{u} \hat{f}_1'(s)\,ds\right) =$$

$$\frac{\sigma^2}{h^2(n)} \int_{1/2}^{u} \int_{1/2}^{u} \sum_{j=0}^{n-1} K\left(\frac{j/n - s}{h(n)}\right) \left\{ 2K\left(\frac{j/n - t}{h(n)}\right) - K\left(\frac{\frac{j-1}{n} - t}{h(n)}\right) - K\left(\frac{\frac{j+1}{n} - t}{h(n)}\right) \right\} ds\,dt$$

where we restrict $\frac{j-1}{n} \geq 0$ and $\frac{j+1}{n} \leq 1$.

Finally,

$$2\,\mathrm{Cov}\left(\int_{1/2}^{u} \hat{f}_1'(s)\,ds,\ \hat{f}_1(1/2)\right) =$$

$$\frac{2 \cdot \sigma^2}{nh^2(n)} \sum_{j=0}^{n} a_j\, K\left(\frac{\frac{j}{n} - \frac{1}{2}}{h(n)}\right) \cdot \left[\int_{1/2}^{u} \left\{ K\left(\frac{\frac{j-1}{n} - s}{h(n)}\right) - K\left(\frac{\frac{j}{n} - s}{h(n)}\right) \right\} ds \right]$$

where $a_j = \begin{cases} 1/2 , & j = 0 \text{ or } n \\ 1 , & \text{otherwise} \end{cases}$

What does this converge to?

$$nh(n) \cdot \mathrm{Var}\left(\hat{f}_1(1/2)\right) \longrightarrow \sigma^2 \int_{-1}^{1} K^2(t)\,dt$$

$$2K\left(\frac{j/n - t}{h(n)}\right) - K\left(\frac{\frac{j-1}{n} - t}{h(n)}\right) - K\left(\frac{\frac{j+1}{n} - t}{h(n)}\right) \doteq \frac{-1}{n^2 h^2(n)} K''\left(\frac{\frac{j-1}{n} - t}{h(n)}\right)$$

We then look at

$$\frac{-\sigma^2}{nh(n)} \Sigma \int_{1/2}^{u} \frac{1}{h(n)} K\left(\frac{j/n - s}{h(n)}\right) ds \cdot \int_{1/2}^{u} \frac{1}{h(n)} K''\left(\frac{\frac{j-1}{n} - t}{h(n)}\right) dt$$

which converges to $2\sigma^2 \int_{-1}^{1} K^2(v)\, dv$

and

$$2 \quad nh(n) \operatorname{Cov}\left(\int_{1/2}^{u} \hat{f}_1'(s)\, ds \,,\; \hat{f}_1(1/2)\right) \to -2\sigma^2 \int_{-1}^{1} K^2(v)\, dv$$

Thus

$$nh(n) \operatorname{Var}\left(\hat{f}_2(u)\right) \to \sigma^2 \int_{-1}^{1} K^2(v)\, dv \quad .$$

5. Asymptotic variance when X is random.

From section 3.3,

$$\hat{rQ}_2(u) \doteq \hat{rQ}_1(u) + \frac{1}{n} \sum_{j=2}^{n} \left(2nh(n)\right)^{-1} \left(\frac{Y_{[j-1:n]} - Y_{[j:n]}}{1/n}\right) \cdot K\left(\frac{\frac{j-1}{n} - u}{h(n)}\right)$$

Thus $\hat{rQ}_2(u) = \hat{rQ}_1(u) + I(u)$

$$\mathrm{Var}\left(\hat{rQ}_2(u)\right) = \mathrm{Var}\left(\hat{rQ}_1(u)\right) + \mathrm{Var}\left(I(u)\right) + 2\,\mathrm{Cov}\left(\hat{rQ}_1(u),\ I(u)\right) .$$

From section 3.1,

$$\mathrm{Var}\left(\hat{rQ}_1(u)\right) \doteq \frac{1}{nh(n)}\, \sigma^2\left(Q(u)\right) \cdot \int_{-1}^{1} K^2(t)\, dt$$

$$\mathrm{Var}\left(I(u)\right) = \frac{1}{4n^2} \cdot \frac{1}{nh(n)}\ \mathrm{Var}\left(m'(X) \mid X = Q(u)\right) \cdot \int_{-1}^{1} K^2(t)\, dt$$

where $m'(x) = \frac{\partial}{\partial x} E[\, Y \mid X = x\,]$

and $2\,\mathrm{Cov}\left(\hat{rQ}_1(u),\ I(u)\right) = \frac{1}{n} \cdot \frac{1}{nh(n)} \cdot C(u)$

It then follows that

$$nh(n)\ \mathrm{Var}\left(\hat{rQ}_2(u)\right) \longrightarrow \sigma^2\left(Q(u)\right) \cdot \int_{-1}^{1} K^2(t)\, dt .$$

To compute the asymptotic variance of $r\hat{Q}_3(\cdot)$, we proceed again by steps as in section 4.2 :

$$\mathrm{Var}\left(r\hat{Q}_3(u)\right) = \mathrm{Var}\left(\int_{1/2}^{u} r\hat{Q}_1{}'(s)\,ds + r\hat{Q}_1(\tfrac{1}{2})\right)$$

$$\mathrm{Var}\left(r\hat{Q}_1(\tfrac{1}{2})\right) \doteq \frac{1}{nh(n)}\,\sigma^2\left(Q(\tfrac{1}{2})\right)\int_{-1}^{1} K^2(t)\,dt$$

$$\mathrm{Var}\left(\int_{1/2}^{u} r\hat{Q}_1'(s)\,ds\right) = \int_{1/2}^{u}\int_{1/2}^{u} \mathrm{Cov}\left(r\hat{Q}_1'(s),\ r\hat{Q}_1'(t)\right) ds\,dt$$

$$\mathrm{Cov}\left(r\hat{Q}_1'(s),\ r\hat{Q}_1'(t)\right) \doteq$$

$$\frac{1}{nh^4(n)}\left\{\int_0^1\int_0^1 (u\wedge v - uv)\,K'\left(\frac{u-s}{h(n)}\right)K'\left(\frac{v-t}{h(n)}\right) du\,Q(u)\,du\,Q(v)\right.$$

$$\left.+\int_0^1 \sigma^2\left(Q(u)\right) K'\left(\frac{u-s}{h(n)}\right) K'\left(\frac{u-t}{h(n)}\right) du\right\} =$$

$$\int_{1/2}^{u}\int_{1/2}^{u} \mathrm{Cov}\left(r\hat{Q}_1'(s),\ r\hat{Q}_1'(t)\right) ds\,dt = \frac{1}{n}\int_{1/2}^{u}\int_{1/2}^{u} C(s,\ t)\,ds\,dt$$

$$+\ \frac{1}{nh}\int_0^1 \sigma^2\left(Q(x)\right)\frac{1}{h(n)}\left\{K\left(\frac{x-u}{h(n)}\right) - K\left(\frac{x-\frac{1}{2}}{h(n)}\right)\right\}^2 dx\ .$$

Finally,

$$2\,\mathrm{Cov}\left(\int_{1/2}^{u} r\hat{Q}_1'(s)\,do,\ r\hat{Q}_1(\tfrac{1}{2})\right) \doteq -\frac{2}{nh(n)}\,\sigma^2 Q(\tfrac{1}{2})\int_{-1}^{1} K^2(t)\,dt + \frac{\text{constant}}{n}$$

Thus,

$$nh(n)\ \mathrm{Var}\left(\hat{rQ}_3(u)\right) \to \sigma^2\left(Q(u)\right) \cdot \int_{-1}^{1} K^2(t)\,dt \quad .$$

6. Preliminary Conclusions

A study of mean integrated squared error done by Melzer (1978) for sample sizes $n = 20, 50, 100$ allows us to conclude that $r\hat{Q}_2(\cdot)$ does not improve on $r\hat{Q}_1(\cdot)$. Also, there is much to be gained by normalizing the estimators so that the weights add up to 1 exactly. This has no effect on our asymptotic results.

The proposed estimator $r\hat{Q}_4(\cdot)$ was abandoned after a few tries on simulated data because of its oscillating behavior.

REFERENCES

Clark, R. M. (1977). "Nonparametric estimation of a smooth regression function," J. Roy. Statist. Soc. B, 39, 107-113.

Melzer, M. (1978). "Empirical study of quantile regression estimators," to be published.

Parzen, E. (1977). "Nonparametric statistical data science: A unified approach based on density estimation and testing for 'white noise'," Technical Report No. 47, Statistical Science Division, SUNY/ Buffalo.

Yang, S. S. (1977). "Linear function of concomitants of order statistics," Technical Report No. 7, Department of Mathematics, Massachusetts Institute of Technology.